Code-Switching and Multilingual ASR JSALT 2022

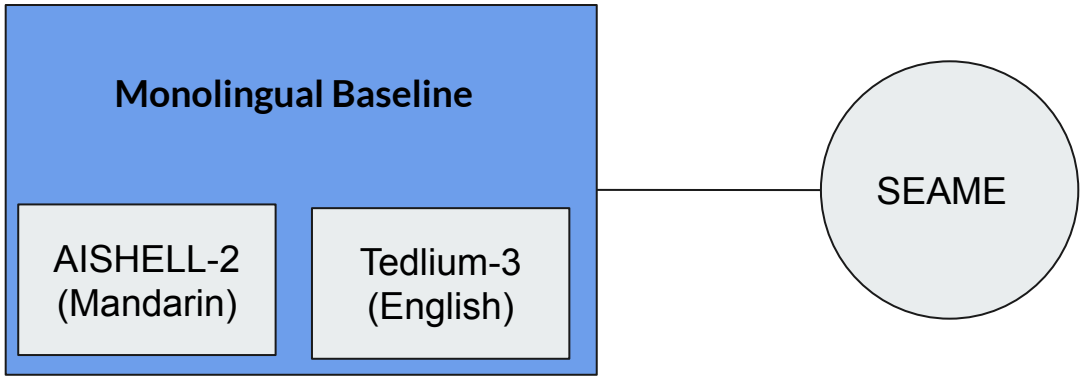
July 6 2022

- WP1 - ASR
- WP2 - CS Generation
- WP3 - Evaluation
- WP4 - Linguistic Aspects of CS

WP1 - ASR

Can we train code-switching ASR systems using only monolingual data?

Mandarin - English ASR



Mandarin - English ASR

	LF-MMI	CTC	RNN-T	EncDec
Tedlium + Aishell	86.4	81.0		
+ Fine-tune 34h mono-lingual SEAME	42.6	55.3		
+ CS LM	27.5	50.1		
Topline trained on all Seame data			22.1	16.7

Mandarin - English ASR

- SEAME corpus is code-switched Mandarin-English
  - Accented English
- Errors due to:
  - missed switches
  - detecting the wrong language
  - accent mismatch

Mandarin - English ASR (Example Errors)

- Wrong language Missed Switch (phonetically similar):

  HYP: the dont turn it jara even the online no quite quite the foreign they also

  REF: 很 多 人 都 在 讲 了 even the online 那 种 怪 怪 的 forum they also

  Hěn duō rén dōu zài jiǎngle Nà zhǒng guài guài de

- Accent ?

  Ref: my mum keeps scold-

  Hyp: my monkey is good

Telugu - English ASR

- Access to small amounts of monolingual Telugu (~50 hours) and larger amounts of Indian-accented English (~150 hours).
- Evaluated on Telugu-English code-switched corpus. 15 hours of CS speech (train) available.
- WER using monolingual Telugu + English + Telugu-English CS speech: 52.3 %
  - Hypothesis contained many instances of intra-word code-switching (E.g., మీడియాEDIA -> MEDIA, PLAలాTFORM -> PLATFORM)
  - Higher fraction of switch points (compared to Mandarin-English CS)

South African ASR

Code-switching is present in low-resource languages, where it might be hard to get transcribed monolingual audio.

Comparison of self-supervised and semi-supervised approaches: labelled English - Xhosa CS data (~3h) + hundreds hours of unlabelled audio data

- Adapting self-supervised pretrained models with LF-MMI
  - Baseline: TDNN + LF-MMI, Pytorch implementation of LF-MMI available in PyChain toolkit
- Standard adaptation with CTC: Hubert as pretrained model + CTC (s3prl): 95.17 without LM

Hyp: BEDELA KUBA SIOFFISINI NOYUKUBA UBENOMKAKGOSE AC YOU

Ref: KUBHETELE UBESE OFFICE KUNOKUBA UBENOMKAKHO OSE I. C. U.

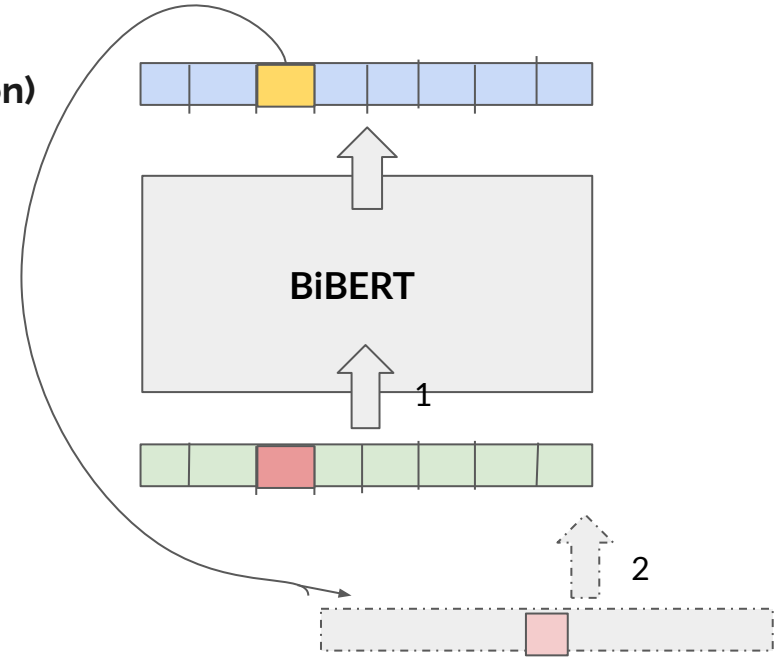
- Extracting features: XLSR-53 features

WP2 - CS Generation

BiBERT (for En-Ar code-switched text generation)

We perform sequential sampling on the BiBERT pretrained on English and Arabic:

a. Start with a code-switched sentence as the seed masked in the first token position.
b. Pass it through the model
c. Decode the predicted masked position in one of the three ways:
    i. Greedy decoding
    ii. Top k (3) decoding
    iii. Top k% (15%) decoding
d. Pass the decoded sequence now masked in the next position back into the model and repeat…

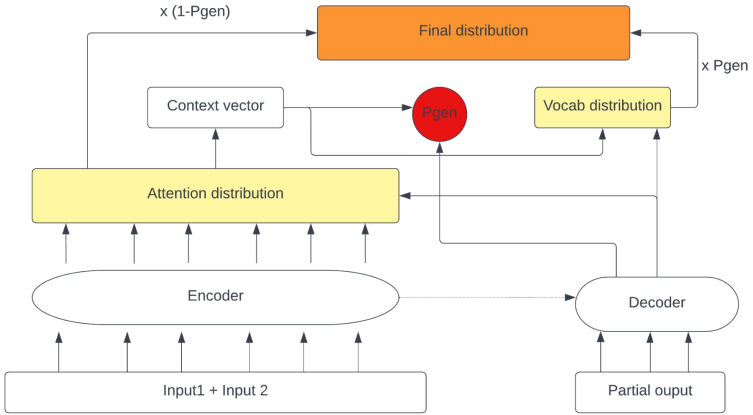


A couple examples…

ليه ليه كت فى دماغك wedding planner ؟	okay تقدرى تعيشى من غير mobile ؟
كت فى دماغك wedding planner ؟	هل تقدرى تعيشى من غير mobile؟
ماذا في حقيبة wedding dress؟	هل تستطيع أن تنمى على a budget؟
ماذا عن ظاهرة black market؟	هل تستطيع أن تحصل yourself into a relationship؟
ماذا عن ظاهرة black الثلج؟	هل يجوز ان throw المال in a puddle؟

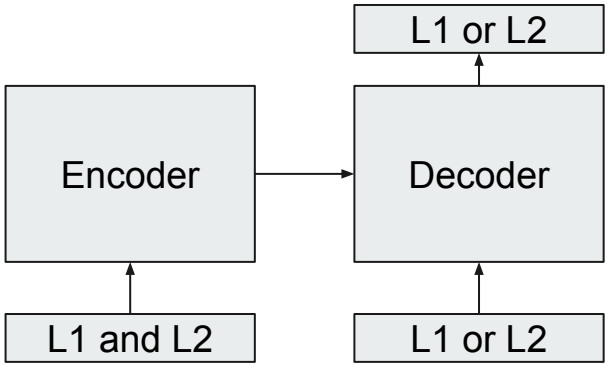
Pointer-Generator Networks

- The model can choose to copy (attention distribution) or generate new words (vocab distribution) from a fixed vocabulary.
- Output is code-switched sentence
- Input1 and input2 can be paired monolingual sentences (L1 and L2)
- Input1 can be monolingual sentence (L1), input2 can be phrase or monolingual sentence containing the to-be-switched phrase from L2



Constrained Decoding

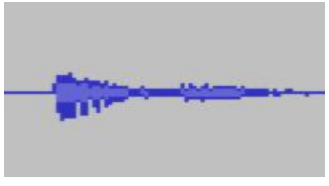
- The model is a (transformer) encoder-decoder model
- Input is two sentences, monolingual L1 and monolingual L2, translations of each other
- Model is trained to output the same L1 sentence half the time and the same L2 sentence half the time
- To generate code-switched output, use grid-beam-search to do constrained decoding among sentences with different number of switch points



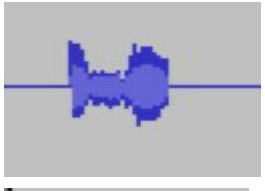
Synthetic Audio Data Generation

We want to generate audio for sentence: “god it حلو you رماك الهوا لا وقعت you”
We use word-based unit-selection with units extracted from monolingual corpora

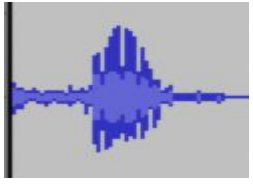
you



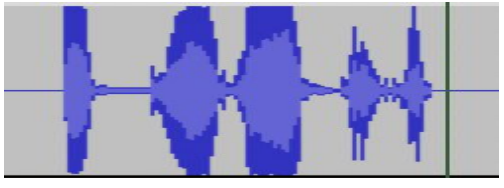
god



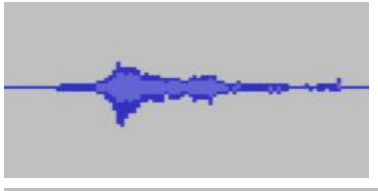
it



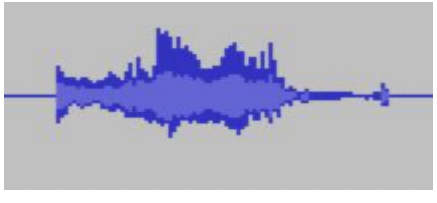
وقعت
vaghaat



حلو
helu

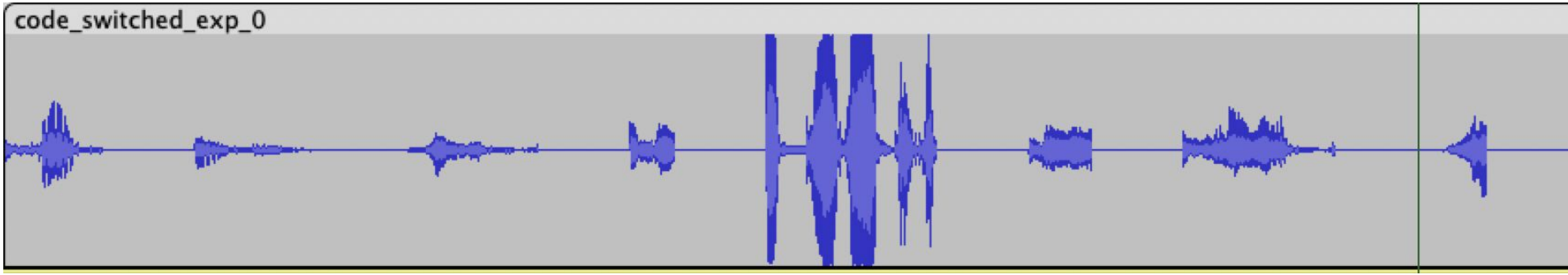


الهوا
alhavaa



Synthetic Audio Generation

These audio segments are spliced together with padding to create code switched audio:



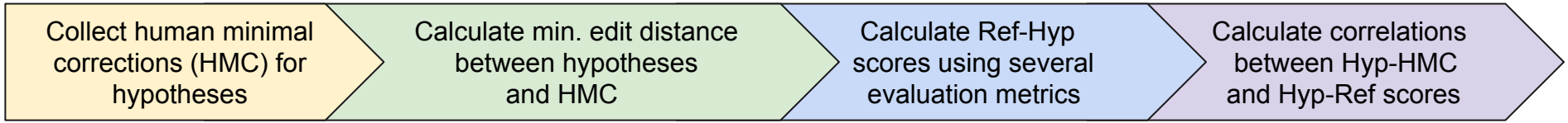
WP3 - Evaluation

WER/CER Example

Ref	ال family mainly weekends فبنزور ال two families و او لو عندنا تمارين برضه او لو عاوزين بقى نتفسح يعنى
Hyp	الويك أند زمايلي فاعملي فا بالنسور families واو لولا عندنا تمارين بوردو أو لو عايزين بقى نتفسح يانغ
Min.Cor.	الويك أندز ماينلي فاميلي فبنزور ال two families واو لو عندنا تمارين بردو أو لو عايزين بقى نتفسح يعني

	Hyp-Ref	Hyp-Min.Cor.
WER	70.0	40.8
CER	47.4	20.0

Overall Plan



- Guidelines for human minimal correction annotation: https://docs.google.com/document/d/1S_LXwfcFR9gDZlJ0V8Pp-7r1dMm3GEen_hsFuxYNfZI/edit
- Use hypotheses from 3 systems; 1 HMM-DNN and 2 E2E
- Annotate 2h of speech (1.3k sentences) X 3 systems
- Evaluation metrics:
  - WER, CER, and MER (Match Error Rate)
  - Transliteration
  - Phone edit distance
- Languages: Egyptian Arabic-English and Telugu-English

Results [Ar-En] - Data and Correlations

- Annotation data:
  - The 3.9K sentences (1.3KX3 systems) are being annotated by 4 Ar-En bilingual annotators.
  - We sampled 200 sentences to be annotated by all annotators for IAA.
    These sentences are already annotated.
  - For the rest of the data, we have 1000/3700 sentences annotated.
- Correlations between Hyp-HMC (CER) and Hyp-Ref Scores (for the 200 sentences):

	CER	WER	MER
Correlation	0.763	0.422	0.503

Results [Ar-En] - Transliteration

	Hyp	Ref
Original	الخبيز بتاعتي اختفت بقى إحنا عارفين كده	ال hobbies بتاعتى اختفت احنا عارفين كده
Tr-En	Alajbez Bettati Akhtift Boca Ahana Arvin Kadeh	al hobbies Bettati Akhtift Ahana Arvin Kadeh
Tr-Ar	الخبيز بتاعتي اختفت بقى إحنا عارفين كده	ال هوبيس بتاعتى اختفت احنا عارفين كده

	Transliteration CER		CER	Transliteration WER		WER
	Tr-En	Tr-Ar		Tr-En	Tr-Ar	
Error Rate	22.1	22.1	27.6	46.1	60.8	61.4

Results [Ar-En] - Phone similarity edit distance

- Map the script from the two languages into IPA phones
- Use the phoneme error rate with substitution weight scaled by the similarity between the phones
- Measure the similarity between the phonemes based on the articulation feature vectors: nasal, front, back, labial etc
- Example:
  - Arabic: ا كايند اف
  - English: `a kind of`
  - Arabic phonetics: `a kajnd aof`
  - English phonetics: `ə kajnd ʌv`
    - PER: 0.5
    - PER_sim: 0.155

		ə	k	a	j	n	d	ʌ	v
a	[	0.	1.	2.	3.	4.	5.	6.	7.	8.],
k	[	1.	0.113	1.113	2.	3.	4.	5.	6.	7.],
a	[	2.	1.113	0.113	1.113	2.113	3.113	4.113	5.113	6.113],
j	[	3.	2.113	1.113	0.113	1.113	2.113	3.113	4.113	5.113],
n	[	4.	3.113	2.113	1.113	0.113	1.113	2.113	3.113	4.113],
d	[	5.	4.113	3.113	2.113	1.113	0.113	1.113	2.113	3.113],
a	[	6.	5.113	4.113	3.113	2.113	1.113	0.113	1.113	2.113],
o	[	7.	6.113	5.113	4.113	3.113	2.113	1.113	0.21	1.21],
f	[	8.	7.081	6.113	5.113	4.113	3.113	2.113	1.21	0.581],
	[	9.	8.081	7.113	6.113	5.113	4.113	3.113	2.21	1.242]]

WP4 - Linguistic Aspects of CS

Are the methods being developed in other work packages generalizable?

A systematic analysis of code-switching across languages and domains

- Ideally, methods developed within the other work packages should be generalizable …
- … but code-switching as a linguistic phenomenon is ill-defined and variable:
  - Amount of code-switching (symmetric or asymmetric)
  - Code-switch points and predictors/triggers of code-switch points
  - Acoustic properties at switch points
- We predict this variability is not random but influenced by factors like:
  - **The language pair (typologies of each language; the linguistic, socio-historic, genealogical relationship between them)**
  - **The domain / context / situation**
  - The speakers (e.g. personality, gender, age of each speaker; the relationship between the speakers)
- Can we identify, systemize, and model this ?

First steps

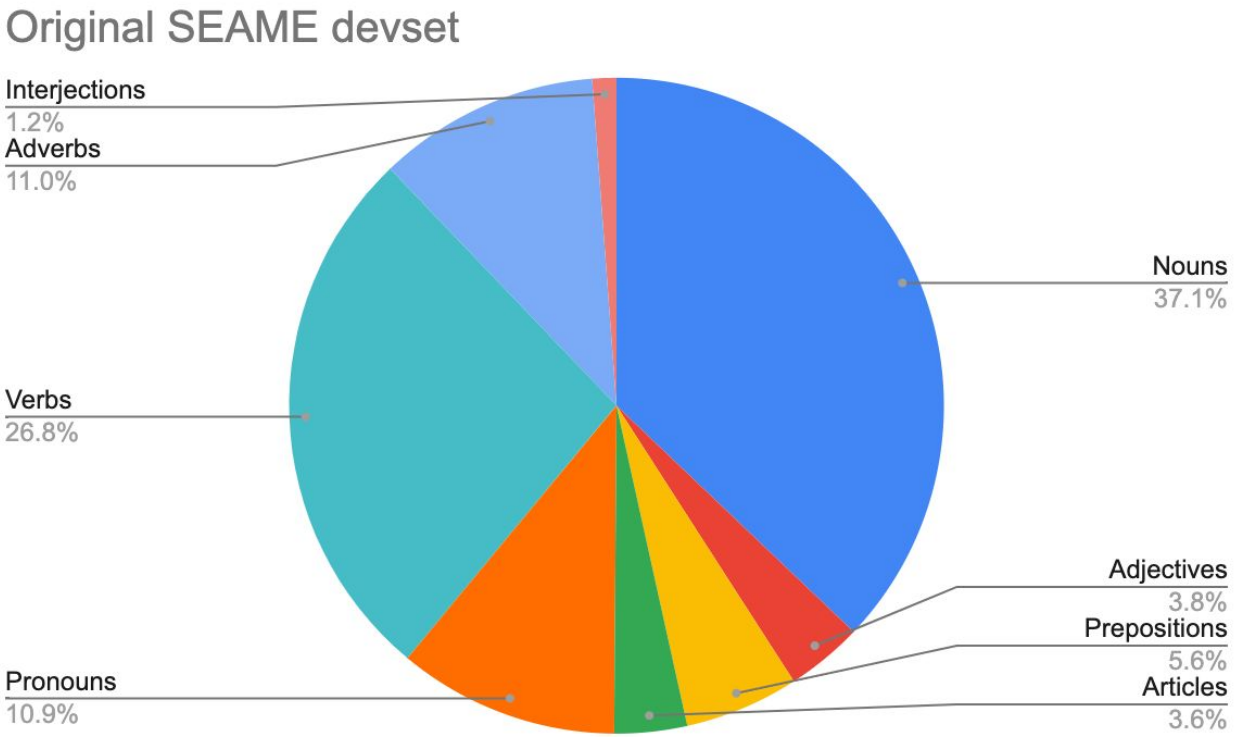
- Collecting data-sets across many different languages and domains, including:
  - Mandarin-English (e.g. SEAME; Datatang)
  - Spanish-English (Bangor Miami)
  - isiXhosa-English (Soap Opera data; self-collected WhatsApp voice notes)
  - Scottish Gaelic-English (audiobooks; web-scraped; MG Alba)
- Defining ‘code-switching richness’ metrics: deciding upon features which can help us identify the ‘richness’ of code-switching in any one data-set
  - Extracting features for baseline variants of this metric, e.g. POS counts, language token counts …
  - Assigning preliminary code-switching richness scores to all data-sets
- Considering variables which may affect, explain, or *predict* this code-switching richness
  - E.g. topic; sociolinguistic, historical, or geographical properties of the language(s); formality
  - Quantifying these variables; Building feature extraction pipelines

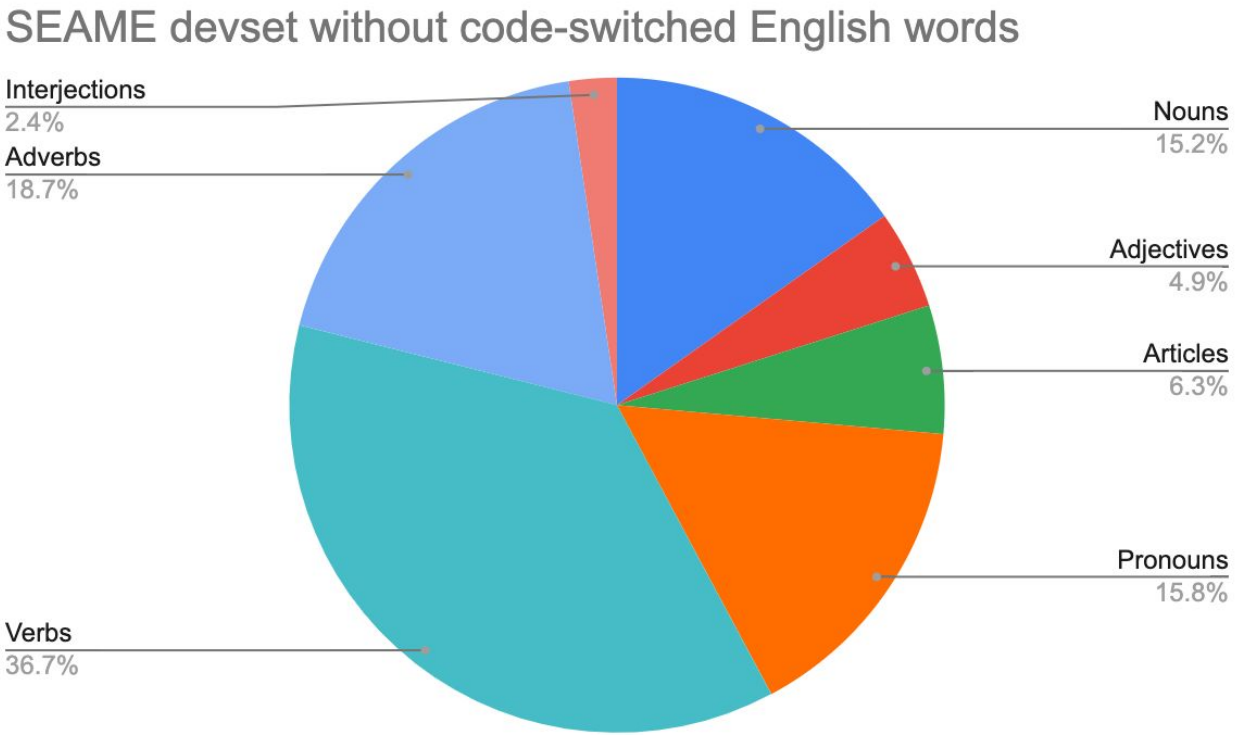
Formality

- Heylighen and Dewaele (1999): *Formality of Language: definition, measurement and behavioral determinants*
  - F-score metric to calculate level of formality in text using distribution of parts of speech

    $$F = (\text{noun frequency} + \text{adjective freq.} + \text{preposition freq.} + \text{article freq.} - \text{pronoun freq.} - \text{verb freq.} - \text{adverb freq.} - \text{interjection freq.} + 100)/2$$

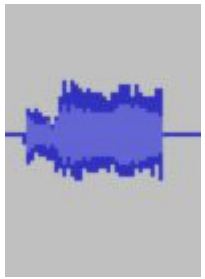
  - \+ POS are correlated with greater formality; - POS are correlated with less formality
- Is this metric reliable?
  - Calibrated on
    - informal Switchboard corpus (F-score = 42% formal)
    - informal TV corpus (F-score = 43% formal)
    - formal broadcast news corpus (F-score = 67% formal)
    - formal legal corpus (F-score = 71% formal)
- SEAME Mandarin-English corpus dev. set
  - F-score = 50% formal on original data
  - F-score = 34% formal excluding code-switched English
- So, Mandarin in code-switching contexts seems to be informal → is this true across languages? We will perform the same analysis on other language corpora.


Original SEAME devset
Interjections
1.2%
Adverbs
11.0%
Nouns
37.1%
Verbs
26.8%
Adjectives
3.8%
Prepositions
5.6%
Pronouns
10.9%
Articles
3.6%


SEAME devset without code-switched English words
Interjections
2.4%
Adverbs
18.7%
Nouns
15.2%
Adjectives
4.9%
Articles
6.3%
Pronouns
15.8%
Verbs
36.7%

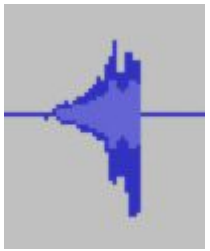
Questions?

Synthetic Audio Data Generation



لا

laa



you

Words not in the corpora used to generate these mappings are skipped..